KazBench-KK: A Cultural-Knowledge Benchmark for Kazakh

Sanzhar Umbet[♡] **Sanzhar Murzakhmetov**[♡†] **Beksultan Sagyndyk**[♡†]
Kirill Yakunin[♡] **Timur Akishev**[♠] **Pavel Zubitskii**[♡]
[♡] Horde Research [†] Yieldmo [♠] KIMEP University
{sanzhar_u,sanzhar}@horderesearch.com

Abstract

We introduce **KazBench-KK**, a comprehensive 7,111-question multiple-choice benchmark designed to assess large language models' understanding of culturally grounded Kazakh knowledge. By combining expert-curated topics with LLM-assisted web mining, we create a diverse dataset spanning 17 culturally salient domains, including pastoral traditions, social hierarchies, and contemporary politics. Beyond evaluation, KazBench-KK serves as a practical tool for field linguists, enabling rapid lexical elicitation, glossing, and topic prioritization. Our benchmarking of various open-source LLMs reveals that reinforcement-tuned models outperform others, but smaller, domain-focused fine-tunes can rival larger models in specific cultural contexts.

1 Introduction

Kazakh reflects a web of pastoral traditions, kinship rules, and post-Soviet social change content that is almost invisible in the English-dominated web. Kazakh is a language that is primarily used and spoken in Kazakhstan and some neighboring regions, but mainstream language models rarely handle it well.

In the NLP landscape, Kazakh is considered a low-resource language due to the scarcity of openly available datasets. This consequently leads to poor performance of LLMs comprehending Kazakh speech and texts, and significantly makes them lack the culturally-specific knowledge of Kazakh traditions, customs and cultural context that are essential for creating inclusive and locally relevant AI systems. While recent efforts have produced datasets for tasks like named entity recognition, sentiment analysis and translation, these are often limited in scope and do not reflect the deep cultural grounding necessary to evaluate how well language models truly understand Kazakh society.

In this paper, we present a semi-automated pipeline designed to generate a benchmark focused on culturally significant knowledge in the Kazakh language. Our approach combines manual topic curation with LLM-assisted keyword generation, automated web retrieval and preprocessing, and context-driven QA generation, followed by both automatic filtering and human validation.

Beyond evaluation, our benchmark opens up practical use cases for linguists working with underrepresented languages. A culturally aware LLM can offer significant advantages to field linguists by connecting language and culture in efficient and innovative ways. Field linguists, who have traditionally relied on the manual collection of linguistic data, can now use LLMs to obtain quick summaries of culture-specific linguistic phenomena and determine which topics are worth further investigation.

Furthermore, both traditional data preparation tasks, including glossing, elicitation prompt construction, and other background research in general and situational decision-making procedures during fieldwork can benefit from these improvements. It is also possible to compare manually collected field data with AI-generated data.

A culturally aware LLM offers field linguists an efficient bridge between language and culture. Instead of relying solely on labor-intensive manual collection, they can query KazBench-KK-tuned models for rapid overviews of culture-specific phenomena, pinpoint promising domains for deeper elicitation, and automatically generate glosses or prompts. Moreover, the benchmark's hierarchical taxonomy reveals how Kazakh speakers organise concepts, turning traditional fieldwork into a more quantified and streamlined endeavour. The accompanying league table allows practitioners to quickly see which publicly available models consistently demonstrate culturally accurate and context-aware responses.

Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics (FieldMatters), pages 38–57

Our contributions are as follows:

- **Introduction of Cultural Benchmark:** We introduce KazBench-KK, a 7111-question multiple-choice benchmark specifically designed to evaluate large language models' understanding of culturally grounded knowledge in the Kazakh language. This benchmark fills a critical gap in resources for evaluating how well AI models understand the nuances of Kazakh culture.

- **Culturally Salient Domain Coverage:** The benchmark covers 17 culturally significant domains, including pastoral traditions, social hierarchies, and contemporary politics. These domains were carefully selected, combining expert-curated topics with LLM-assisted web mining, ensuring a comprehensive and relevant assessment of cultural understanding.

- **Semi-Automated Pipeline for Data Generation:** We present a novel, semi-automated pipeline for the efficient generation of high-quality, culturally relevant data. This pipeline combines the strengths of both human expertise and machine automation, addressing the challenges of data scarcity for low-resource languages.

- **Benchmarking of Open-Source LLMs:** The paper includes a thorough benchmarking of several open-source large language models. This provides a valuable resource for linguists and practitioners seeking to choose the most appropriate models for tasks that involve the Kazakh language and its cultural context.

2 Related Work

Prior work on evaluating cultural knowledge falls into three strands: general English benchmarks, multilingual suits, and recent Kazakh-specific sets. They effortlessly handle multiple languages, generate text with human-like fluency, and are useful in many contexts. However, despite their global reach, these models remain heavily "westernized", and predominantly understand and reflect Western cultural norms and traditions (Naous et al., 2024; Wang et al., 2024; Cao et al., 2023). This western-centric bias inevitably creates a gap when it comes to accurately interpreting and engaging with non-Western, particularly Central Asian, cultures.

Multiple studies have analyzed the performance of language models to generate culturally relevant responses in diverse cultural settings. However, most of these evaluations are centered around high-resource languages, or rely mainly on translation-based approaches that fail to capture deep cultural context. To situate our work, we first review existing English language benchmarks, then discuss recent efforts to extend such benchmarks to multilingual or indigenous settings. Finally, we highlight the current limitations of Kazakh language resources and demonstrate how our work addresses this critical gap.

2.1 General-purpose English Benchmarks

Currently, there are multiple benchmarks in English that try to assess models' different aspects of knowledge. For example, the general language understanding evaluation (GLUE; Wang et al., 2018) and SuperGlue (Wang et al., 2019) benchmarks are aimed to evaluate language models on multiple tasks, including: sentiment analysis, lexical entailment, coordination scope and many more. Moreover, HellaSwag (Zellers et al., 2019) and CosmoQA (Huang et al., 2019) benchmarks are also commonly used to evaluate commonsense reasoning. Nevertheless, as the development of language models progress, it became more common for them to perform on these benchmarks on the human-like level. Therefore, to make better assessments of more advanced language models new challenging benchmarks were developed. They include: MMLU (Hendrycks et al., 2021b,a), AGIEval (Zhong et al., 2023) and BIG-bench (Srivastava et al., 2022), each introducing more complex questions on different topics.

2.2 Multilingual & Cross-cultural Benchmarks

The evaluation of LLMs across different languages has led to the creation of several multilingual benchmarks. Notable examples include XGLUE (Liang et al., 2020), XTREME (Hu et al., 2020), and MEGA (Ahuja et al., 2023), which are designed to test language models' performance on a range of tasks in multiple languages, from high-resource to low-resource ones. Additionally, efforts have been made to build datasets tailored to specific language families (Huang et al., 2023; Doddapaneni et al., 2023; Adebara et al., 2023).

These benchmarks mainly assess syntactic and semantic capabilities such as translation, question answering, and classification.

Beyond general linguistic evaluation, more recent research has focused on cultural benchmarks that aim to measure LLMs' understanding of sociocultural knowledge. These include datasets like GeoLAMA (Yin et al., 2022), which evaluates geo-diverse commonsense reasoning, and CulturalAtlas (Fung et al., 2024), which compiles social norms from over 193 countries. Other works, such as CREHate (Lee et al., 2024) and StereoKG (Deshpande et al., 2022), examine cultural stereotypes and bias across regions using social media and crowd-sourced data. However, none of these suits addresses the cultural fabric of Kazakh life.

2.3 Kazakh-specific Benchmarks

Despite recent advancements in multilingual NLP, Kazakh remains significantly underrepresented in benchmark development. While foundational datasets have been introduced for core NLP tasks, such as KazNERD for named entity recognition (Yeshpanov et al., 2022), KazSAnDRA for sentiment analysis (Yeshpanov and Varol, 2024), and KazParC for machine translation (Yeshpanov et al., 2024) - most of these are narrow in scope and task-specific. They offer valuable building blocks, but do not capture the broader reasoning capabilities or cultural depth needed to evaluate how well LLMs understand Kazakh society.

To help address this, a few benchmark-style datasets have recently emerged. One example is the Kazakh Unified National Testing MC dataset, which contains nearly 15,000 multiple-choice questions pulled from Kazakhstan's national standardized exams (Sagyndyk et al., 2024b). These questions span subjects such as Kazakh literature, history, geography, and biology, providing a realistic and academically grounded way to test the grasp of a model of school-level Kazakh knowledge.

Another effort is the Kazakh Constitution MC dataset, which includes more than 400 multiple-choice questions based on Kazakhstan's constitution (Sagyndyk et al., 2024a). This benchmark is more civic in nature, offering a way to evaluate how well a model understands the legal and governmental concepts that are specific to Kazakhstan.

There is also a Kazakh-translated version of the popular MMLU benchmark, containing around 15,900 multiple-choice questions across a wide range of topics (Sagyndyk et al., 2024c). While helpful for assessing general reasoning in a low-resource setting, this benchmark is entirely translation-based and may not fully preserve Kazakh-specific cultural or contextual nuances.

From a field-linguist perspective, an LLM that handles such culturally grounded content could accelerate tasks like domain word-list expansion or contextual translation checks. However, no public benchmarks let practitioners compare models on these abilities.

All of these benchmarks represent important steps forward. But they still focus mostly on academic or formal domains, and none are designed to test a model's ability to reason about everyday Kazakh customs, values or culturally embedded practices. In other words, we still do not know how well LLMs can engage with the lived experience of Kazakh speakers.

3 Methods

The creation of culturally aware NLP models requires considerable effort, particularly for low-resource languages, where even regular data is limited. Data acquisition methods generally fall into three categories, manual, automatic, and semi-automatic (Liu et al., 2025). Manual data acquisition involves hiring native speakers or professional translators to annotate or culturally adapt textual resources. Additionally, crowdsourcing platforms, university mailing lists, and Slack or Discord channels of relevant organizations regularly serve as sources for gathering culturally rich textual data through user interaction, conversations, and public messaging (Liu et al., 2021).

Another promising method for data collection leverages LLMs to extract cultural knowledge. For instance, Nguyen et al. (2023) proposes a workflow that identifies culturally significant information in texts by using named entity recognition, culturally trained classification models, and information retrieval and ranking algorithms to create culturally aware datasets. However, as highlighted by Putri et al. (2024), fully automating dataset creation using LLMs remains challenging, as the generated texts typically lack deep cultural understanding and may exhibit fluency errors. A potential solution to balance automation and quality is to adapt a semi-automatic approach, merging manual annotations with automated processes. Studies by

Liu et al. (2024) and Bhutani et al. (2024) demonstrated the effectiveness of using prompting techniques for initial data generation, followed by human evaluation to verify and refine cultural relevance.

To address the scarcity of culturally grounded Kazakh benchmarks, we developed a semi-automated data generation pipeline that uses LLMs and web-scale retrieval to synthesize high-quality data. The core goal of the system is to generate multiple choice questions centered on culturally and contextually significant topics in Kazakhstan, which are currently absent from existing benchmarks.

3.1 Linguistic & cultural categories

Our selection of categories and concepts was guided by the goal of capturing Kazakh culture in various forms of its representation. We primarily focused on those aspects of culture that can be expressed, preserved, or transmitted through language and text, whether spoken or written. The inherently textual categories that we added to the dataset are related to (1) creativity (literature, song lyrics, and films) and (2) formulaic language (proverbs, sayings, prayers, and spiritual expressions). Other categories selected for the dataset were not inherently textual in nature, but have been recorded and can be described using text: (3) traditions and customs, as they form the core of any culture, (4) social relations and hierarchies, as they reflect the organization of the society, (5) daily life (names of traditional foods and clothing and terminology used to refer to traditional household objects, architecture, and agriculture), and (6) arts and crafts (tools, materials, and techniques).

Criteria	**Description**
Traditions	Family events; holidays, rituals and ceremonies
History	Crucial historical events; historical figures
Social relationships	Family members; relatives; polite terms for strangers; endearments for loved ones
Politics and social strata	Historical terms (e.g., khans, bis); zhuzes and rus
Proverbs, spirituality	Sayings, spiritual terms (e.g., *bata*); superstitions, mythology
Humor	Jokes, *aitys*, humorous figures (e.g., Aldar Kose); wordplay
Cuisine	Recipes; names for food and beverages
Sports and games	Names and rules of traditional games and sports
Films	Classic and contemporary Kazakh cinema; landmark films, directors, actors, and culturally significant storylines
Literature	Poetry and fiction with cultural relevance
Song lyrics	Traditional songs, *kuys*
Instruments	Names of instruments and parts
Arts and crafts	Crafts, decorative and performing arts
Clothing	Names of traditional garments
Named entities	Names of people/places and their meanings (onomastics)
Agriculture	Terms related to farming and herding
Architecture	Yurt structure and home elements

Table 1: Cultural Knowledge Categories

3.2 Semi-Supervised Benchmark-generation pipeline

Our data generation pipeline consists of several key stages

Topic initialization. Initially, we manually curated a comprehensive list of general topics, organizing them into clearly defined knowledge categories relevant to Kazakh society, such as: Media, Politics, Traditions, and so on. Within each general category, we further identified distinct subcategories to cover diverse perspectives and deepen contextual relevance. For instance, under 'Current social life', we explored subcategories like the scandalous 'Bishimbayev case', ecological issues in Almaty or negligence in the Thermal Plant in Ekibastuz.

LLM-based keyword generation. For each category–subcategory pair, our linguists and sociologists first compiled a concise seed list of culturally salient terms. We then used *GPT-4o* to expand these expert-provided seeds, instructing the models to propose roughly ten additional, culturally anchored keywords (i.e., sub-subcategories) that captured dialectal variation, idiomatic usage, and other nuanced linguistic forms. This human-in-the-loop procedure ensured that domain knowledge grounded the process while the LLM broadened the lexical scope. The resulting keyword sets were subsequently transformed into natural-language search queries, reflecting how a native speaker might phrase them in a typical Google search.

Algorithm 1: KazBench-KK data-generation pipeline

Input: Manually curated category list C with seed keywords
Output: Multiple-choice question set $\mathcal{Q}$

foreach $(c, \mathit{sub}) \in C$ **do**
  `/* Step 1: keyword expansion */`
  $\mathit{Seeds} \leftarrow$ linguist/sociologist seed list ;
  $\mathit{Expanded} \leftarrow$ LLM_Expand($\mathit{Seeds}, n{=}10$) ;
  $\mathit{Queries} \leftarrow$ MakeQueries($\mathit{Seeds} \cup \mathit{Expanded}$) ;
  `/* Step 2: content retrieval */`
  $\mathit{Docs} \leftarrow$ WebSearch($\mathit{Queries}$) ;
  `/* Step 3: preprocessing */`
  $\mathit{Clean} \leftarrow$ ParseAndClean(Docs) ;
  $\mathit{Corpus} \leftarrow$ Deduplicate(Clean) ;
  `/* Step 4: MCQ generation */`
  foreach $d \in \mathit{Corpus}$ **do**
    $\mathit{mcq} \leftarrow$ LLM_MCQ(d) ;
    if IsCultureSpecific*(mcq)* **then**
      $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{\mathit{mcq}\}$;
return $\mathcal{Q}$

Content retrieval. With the search queries generated, we then performed automated web retrieval. We integrated external API services to execute extensive searches on websites and platforms such as Wikipedia, local Kazakh news outlets, and blog posts.

Webparsing and Preprocessing. The retrieved website URLs underwent an automated custom parsing and clearing process. We utilized the open-sourced HTML parsing scripts to scrape textual data from the websites, and implemented preprocessing techniques to remove HTML tags, navigation elements, and redundant information. Additionally, we employed a deduplication approach to ensure data quality and consistency.

LLM-based question generation. After preprocessing, the cleaned text corpus was fed into a large language model to generate structured multiple-choice questions (MCQ). For each content chunk, the LLM was prompted to produce context-based MCQs along with four answer options, with three being distractors and one correct answer, grounded in the specific cultural or historical context. We adopted a four-option format to align with common standardized practices in Kazakhstan and global MCQ benchmarks, ensuring compatibility with existing evaluation tools. To support better dataset usability and analysis, each question was also tagged with a binary annotation indicating whether it required context-specific knowledge, and whether a generated question was Kazakh-culture-specific. This allowed us to later filter and categorize the dataset based on its cultural relevance and reasoning complexity.

3.3 Data Filtering

We developed a set of criteria to ensure the high quality of our data. These criteria applied to both the questions and the answer options, focusing on their overall structure, logic, coherence, grammatical correctness, and the relevance of the options to the questions. We aimed to avoid absurd or overly obvious items and ensure that the answer options, including distractors, were appropriate and justifiable. Additionally, we wanted our data to be balanced in terms of general quality, difficulty, and diversity. Finally, we evaluated the overall relevance of the question–answer pairs to the categories and subcategories constituting the notion of culture. Applying these criteria helped us refine the dataset and eliminate any major illogical, incoherent, absurd, or otherwise irrelevant items.

3.3.1 Automated pre-filtering

To reduce annotator load, we translated the above rules into a binary "keep vs. discard" classifier implemented as a `gemini-2.0-flash-lite` agent in LangChain. The model embeds each MCQ with its answer set, applies chain-of-thought self-critique, and filters out items whose risk score exceeds 0.5 prior to human review. Table 2 presents the classifier's performance on a held-out set of 97 examples; the macro F_1-score is 0.87.

Class	Precision	Recall	F_1	Support
Discard (noise)	0.84	0.88	0.86	42
Retain (good)	0.91	0.87	0.89	55
Accuracy			0.88	97
Macro avg	0.87	0.88	0.87	97
Weighted avg	0.88	0.88	0.88	97

Table 2: Metrics for the binary filter

3.3.2 Human curation

To complement the automatic filter, we collaborated with four native-speaker linguists who manually reviewed and refined the remaining items. Following the same rubric used by the automated filtering agent, the annotators could also correct the wording, swap distractors, or flag entire MCQs for removal; no overlapping assignments or majority voting was required.

Annotator profile. All four annotators are Kazakh women of Asian ethnicity. Three are aged 18–24, and one falls within the 35–44 age range. Two hold undergraduate degrees in Language studies, while the other two have completed master's programs. As a qualification check, each annotator answered ten control questions from Kazakhstan's national standardized exams (Sagyndyk et al., 2024b) for Kazakh language and all scored a perfect 10/10.

ID	Gender	Age	Education	Ethnicity / Nationality
A1	Woman	18–24	B.A. Linguistics	Asian / Kazakh
A2	Woman	18–24	B.A. Linguistics	Asian / Kazakh
A3	Woman	18–24	M.A.	Asian / Kazakh
A4	Woman	35–44	M.A.	Asian / Kazakh

Table 3: Demographic profile of human annotators.

4 Dataset Description

4.1 Overview and format

Statistic	A	B	C	D	question
Tokens (total)	22 425	25 056	24 045	22 193	63 059
Tokens (avg.)	3.154	3.524	3.381	3.121	8.868
Unique tokens	8 997	10 565	10 048	9 297	15 282
Sentences (avg.)	1.009	1.011	1.010	1.009	1.013
Kk-char ratio	0.0907	0.0906	0.0895	0.0874	0.1020

Table 4: Descriptive statistics for answer options (A–D) and question stems (Q).

KazBench–KK consists of **7,111** multiple-choice questions (MCQs).[1] Each JSON record contains a single-sentence stem in Cyrillic Kazakh, four answer options (`A-D`), a field indicating the correct answer, and three metadata fields (`category`, `subcategory`, `keyword`).

4.2 Quantitative characteristics

Category distribution. Figure 2 shows that cultural topics are highly uneven on the web and the dataset mirrors this reality: *History* is the largest class with *1 103* items, followed by *Onomatopoeia* (621) and *Agriculture* (579). The smallest bar belongs to *Swearing* category with slightly over 50 questions. Despite the long tail, every category contains dozens of samples, enabling per-domain evaluation.

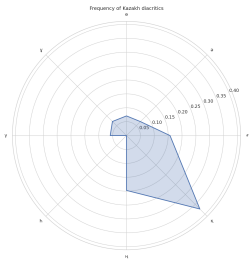

Figure 1: Diacritics distribution

Sub-category coverage. The finer-grained view (Fig. 3) contains 70-plus sub-categories. Counts range from roughly 450 questions at the top to around 30 at the bottom, implying that no single niche dominates the benchmark.

Question length. Box plots in Fig. 4 reveal a tight span: the median stem length is **7 tokens** across all domains, with the middle 50 % of examples falling between 6 and 9 tokens. Only a handful of outliers exceed 14 tokens.

Lexical diversity. Token-type ratios by column are plotted in Fig. 6. Stems have the lowest variety, reflecting repeated use of interrogatives (*қандай, қай*). Answer options are markedly richer (TTR $\approx$ 0.60–0.70), and some specialised domains (e.g. *Swearing expressions*) push the ratio beyond 0.95.

Orthographic coverage. Eight Kazakh-specific Cyrillic letters (*ә, қ, ң, ғ, ө, ү, ұ, і*) appear in the corpus. The radar chart (Fig. 1) shows that "*қ*" alone accounts for about 40 % of the diacritic tokens, with "*ң*" and "*ғ*" the next most common. Consequently, automatic evaluation cannot succeed by handling only Russian spellings.

Answer-key balance. The answer keys were originally placed so that each position (A-D) had the correct option exactly one quarter of the time, eliminating positional bias at generation time. After human curation, where annotators occasionally rewrote, swapped, or pruned options, the distribution drifted, and Fig. 7 now shows a modest skew across positions. We report this shift to inform reviewers about the residual position bias introduced during manual cleanup.

[1] Available at HF.

4.3 Linguistic profile

Frequent vocabulary. The histogram in Fig. 5 confirms that stems are dominated by function words and wh-terms, whereas answers introduce content words such as *қазақ* 'Kazakh', *дәстүрлі* 'traditional' and named entities. This design forces models to rely on content-specific cues rather than stereotyped question templates.

Category–specific variation. The heat-map of type–token ratios highlights clear lexical contrasts: creative domains such as *Cinema* display the highest diversity within columns, while everyday areas (*Agriculture*, *Traditions*) use a narrower but still non-trivial vocabulary. Such variation allows error analysis that links model failures to specific cultural sublexica.

Summary. Taken together, the figures demonstrate that KazBench–KK offers (i) broad topical coverage, (ii) compact but information rich stems, (iii) balanced answer positions, and (iv) authentic Kazakh orthography. These properties make the dataset a realistic stress test for language models that claim cultural knowledge of Kazakhstan.

5 Results

We selected a diverse panel of 21 checkpoints that (i) span the major open-source families (Llama-3, Gemma-3, Qwen 2.5, Mistral, Nemotron, DeepSeek) and (ii) cover the full spectrum of tuning regimes (base SFT, community SFT-tune, and RL/Instruct). We excluded any model that participated in our data-generation pipeline—those very large, API-only LLMs that seeded the MCQs-because evaluating them on a benchmark they helped create would inflate scores and mask true generalisation. This "no-leak" policy avoids circularity and lets us gauge how well *independent* models, with parameter counts from 8B to 70B, handle culture-specific content. Within that cohort, reinforcement-/instruction-tuned models dominate

On logit-level multiple-choice scoring, reinforcement-/instruction-tuned models dominate: Gemma-3-27B-*it* (0.72), both Llama-3-70B Instruct variants (0̃.71), and Nvidia's Nemotron-Super-49B RL model (0.69) form a clear first tier. Model scale still matters - Nemotron-Nano-8B RL plunges to 0.35 - but domain-focused fine-tunes can partly offset size: the 8B Sherkala chat model (0.69) and KazLLM-70B (0.69) rival much larger base checkpoints. Pure SFT baselines such as Gemma-3-12B-pt (0.62) and Qwen-32B (0.62) trail their RL counterparts by 6–10 points, confirming the benefit of preference optimization even when no text generation is required. Overall, reinforcement alignment combined with sufficient parameters remains the most reliable recipe for KazBench-KK accuracy, though well-targeted community SFTs can yield competitive gains.

At the category level, *Cinema* and *Onomatopoeia* are consistently the hardest sections, dipping below 0.60 for nearly every model, including top-tier Gemma-3-27B-*it* (0.69 and 0.67, respectively) and falling into the mid-0.40s for smaller checkpoints. Conversely, politically grounded knowledge is easy: all first-tier models top 0.79 on *Politics & Social Stratification*, with Gemma-3-27B-*it* at 0.79 and Llama-3-70B Instruct at 0.81. Nvidia's Nemotron-Super-49B shows a distinctive strength in *Musical Instruments* (0.69) and *Architecture* (0.72), whereas the Sherkala 8B chat model punches above its weight in *Humor* (0.71) and *Cuisine* (0.67)-categories where many SFT baselines lag. KazLLM-70B peaks at *Swearing & Offensive Expressions* (0.70), reflecting its culture-specific tuning. The overall spread suggests that cultural trivia tied to media, sound symbolism, and pop-culture films remains challenging, while hierarchical or historically codified knowledge (political titles, social classes, formal rituals) is much easier for models to retrieve.

Model name	Type	Accuracy
google/gemma-3-27b-it	rl	0.7216
meta-llama/Llama-3.3-70B-Instruct	rl	0.7090
meta-llama/Llama-3.1-70B-Instruct	rl	0.7030
nvidia/Llama-3_3-Nemotron-Super-49B-v1	rl	0.6936
inceptionai/Llama-3.1-Sherkala-8B-Chat	sft-tune	0.6909
issai/LLama-3.1-KazLLM-1.0-70B	sft-tune	0.6892
google/gemma-3-12b-it	rl	0.6794
mistralai/Mistral-Small-24B-Instruct-2501	rl	0.6761
Qwen/Qwen2.5-32B-Instruct	rl	0.6334
google/gemma-3-12b-pt	sft	0.6241
Qwen/QwQ-32B	sft	0.6165
deepseek-ai/DeepSeek-R1-Distill-Llama-70B	sft	0.6019
Qwen/Qwen2.5-14B-Instruct	rl	0.6002
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	sft	0.5996
google/gemma-3-4b-pt	sft	0.5854
google/gemma-3-4b-it	rl	0.5828
meta-llama/Llama-3.1-8B-Instruct	rl	0.5750
issai/LLama-3.1-KazLLM-1.0-8B	sft-tune	0.5656
nvidia/Llama-3.1-Nemotron-Nano-8B-v1	rl	0.3542
TilQazyna/llama-kaz-instruct-8B-1	rl	0.2768

Table 5: Overall accuracy of evaluated models. Model types: **rl** = reinforcement-tuned, **sft** = base supervised fine-tune, **sft-tune** = post supervised fine-tune.

Why an Offline-Only Evaluation All checkpoints were executed locally-without any hosted-API calls-for four technical reasons.

(1) Apples-to-apples comparability: restricting the pool to models that ship raw weights prevents API-only systems from benefiting from undisclosed tool use or server-side retrieval, so every score reflects the base language model alone.

(2) Decoding transparency: local inference lets us pin the exact tokenizer build, sampling algorithm, and context window; commercial endpoints may apply proprietary post-processing that we cannot inspect or replicate.

(3) Logit access for analysis: computing per-option log-likelihoods, error heat-maps, or calibration curves requires raw logits-information that most APIs do not expose.

These constraints keep the leaderboard a clean test of model weights, tokenization, and decoding policy-nothing else.

6 Conclusions

This study introduces *KazBench-KK*, a 7,111-item benchmark that assesses how well contemporary language models grasp cultural knowledge encoded in Kazakh. Built through a semi-automatic pipeline that blends expert guidance, web mining, and careful human curation, the dataset covers seventeen domains ranging from clan hierarchy to popular cinema.

The evaluation paints a mixed picture. Large, reinforcement-aligned models, like Gemma-3-27B-it and the Llama-3-70B Instruct pair-handle codified facts such as historical events with confidence, but their accuracy drops on items tied to film references or sound-symbolic words. Smaller community fine-tunes, notably Sherkala-8B and KazLLM-70B, narrow the gap in conversational categories like humour, swearing, and cuisine, showing that targeted data can offset limited parameter count in specific niches.

Practically, the league table offers a guide: Choose a heavyweight model when the task demands institutional knowledge, and reach for a lean, locally tuned model when nuance in everyday language matters more. For researchers, the consistent underperformance on Cinema and Onomatopoeia highlights clear gaps where additional data collection is likely to yield rapid gains.

Finally, the methodology itself is portable. Because each stage of the pipeline-seed selection, keyword expansion, retrieval, and filtering-relies on general tools, other language communities can replicate the process to create their own culturally specific benchmarks.

7 Future Work

Future research could expand KazBench-KK by integrating open-ended questions and dialect-specific knowledge from underrepresented rural regions. Moreover, the semi-automated benchmarking pipeline introduced here can be extended beyond textual data, facilitating culturally grounded benchmarks in multimodal domains such as images, audio, and video. Applying this methodology across diverse modalities would support a more comprehensive understanding and representation of Kazakh culture and other low-resource cultural contexts.

8 Limitations

Our benchmark cannot claim exhaustive coverage of Kazakh culture. Web-derived material is skewed toward urban, Russian-influenced outlets, so the lexicon of rural dialects and oral genres (e.g., regional *aitys*) remains underrepresented. Although the generation pipeline balanced answer keys at creation time, manual curation introduced a mild positional skew (Fig. 7). The questions are single-sentence MCQs; they do not test open-ended generation, discourse planning, or code-switching.

9 Ethics

Data provenance. All text was scraped from publicly accessible websites; we removed pages that contained personal names, contact details, or paywalled material. The released dataset stores only short question stems and answer options, minimising potential copyright concerns.

Annotator welfare. Four native-speaker linguists contributed on a *voluntary* basis; they received no monetary compensation, but gave their informed consent, could skip any item, and were free to withdraw at any time.

Bias and cultural sensitivity. Web sources may reflect gender, regional, or political biases; the benchmark therefore inherits those biases. Some items reference sensitive topics (e.g. clan affiliation, swearing); we flagged such questions with metadata so that downstream users can filter them if desired.

Acknowledgments

We thank Arman Zharmagambetov for his valuable feedback and insightful discussions that significantly contributed to the development of this work.

We also thank the reviewers for their thoughtful comments and suggestions, which helped improve the quality and clarity of the paper.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67, Dubrovnik, Croatia. Association for Computational Linguistics.

Awantee Deshpande, Dana Ruiter, Marius Mosbach, and Dietrich Klakow. 2022. StereoKG: Data-driven knowledge graph construction for cultural knowledge and stereotypes.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition lm benchmarking.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, jiayi lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Nayeon Lee, Chani Jung, Junho Myung, Jiho Jin, Jose Camacho-Collados, Juho Kim, and Alice Oh. 2024. Exploring cross-cultural differences in english hate speech annotations: From dataset construction to analysis.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv*, abs/2004.01401.

Chen Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of*

the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Tuan-Phong Nguyen, Simon Razniewski, Aparna Varde, and Gerhard Weikum. 2023. Extracting cultural commonsense knowledge at scale. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1907–1917, New York, NY, USA. Association for Computing Machinery.

Rifki Afina Putri, Faiz Ghifari Haznitrama, Dea Adhista, and Alice Oh. 2024. Can LLM generate culturally relevant commonsense QA data? case study in Indonesian and Sundanese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20571–20590, Miami, Florida, USA. Association for Computational Linguistics.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, Sanzhar Umbet, and Kirill Yakunin. 2024a. Kazakh constitution: Multiple choice benchmark. Available on Hugging Face.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, Sanzhar Umbet, and Kirill Yakunin. 2024b. Kazakh unified national testing: Multiple choice benchmark. Available on Hugging Face.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, Sanzhar Umbet, and Kirill Yakunin. 2024c. Mmlu on kazakh language: Translated mmlu benchmark. Available on Hugging Face.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds,

Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint 1905.00537*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. Kaznerd: Kazakh named entity recognition dataset.

Rustem Yeshpanov, Alina Polonskaya, and Huseyin Atakan Varol. 2024. Kazparc: Kazakh parallel corpus for machine translation.

Rustem Yeshpanov and Huseyin Atakan Varol. 2024. Kazsandra: Kazakh sentiment analysis dataset of reviews and attitudes.

Da Yin, Hritik Bansal, Masoud Monajatipoor, Liunian Harold Li, and Kai-Wei Chang. 2022. GeoMLAMA: Geo-diverse commonsense probing on multilingual pre-trained language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2039–2055, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models.

Appendix

A Question Category Distribution

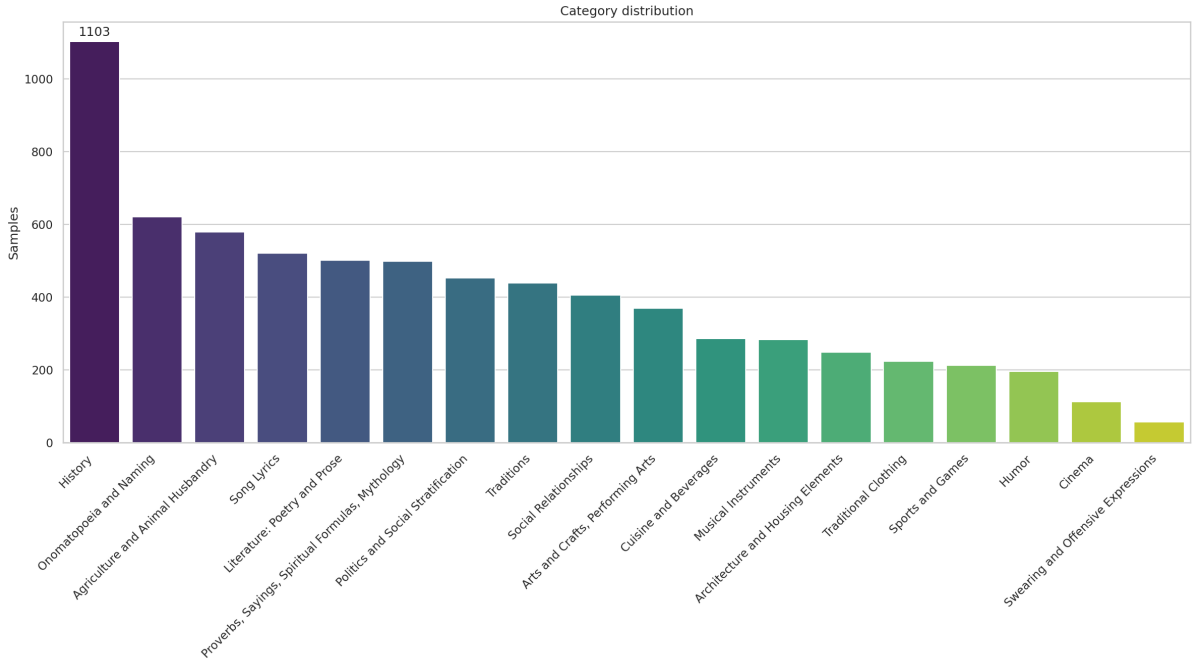

Figure 2: Distribution of questions across major cultural categories in KazBench-KK.

B Sub-Category Distribution

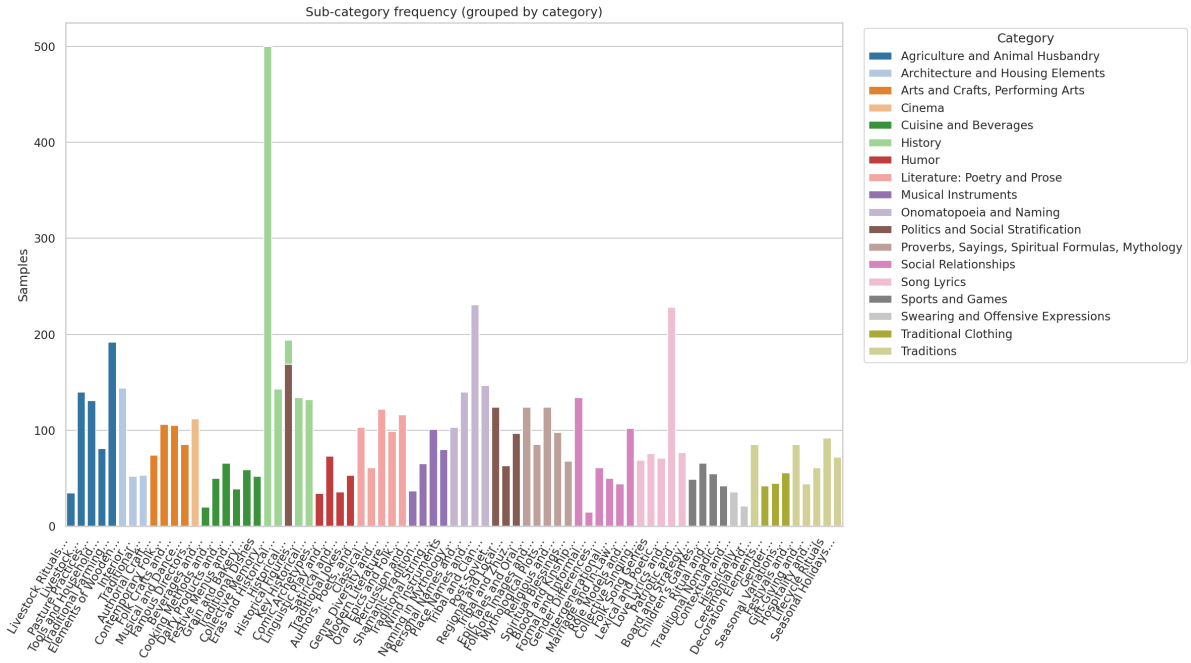

Figure 3: Granular breakdown of question counts per sub-category, demonstrating the breadth of domain-specific coverage.

C Question Length Analysis

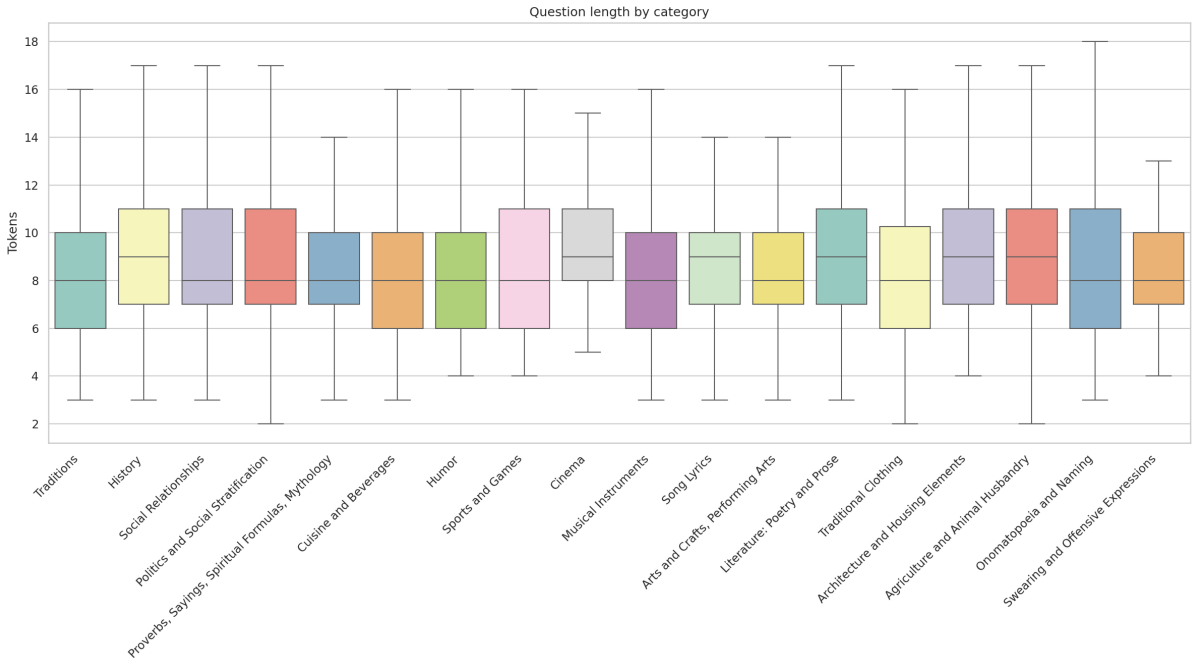

Figure 4: Box plot of question stem lengths (in tokens), showing central tendency and variability across domains.

D Top Token Frequency in Questions

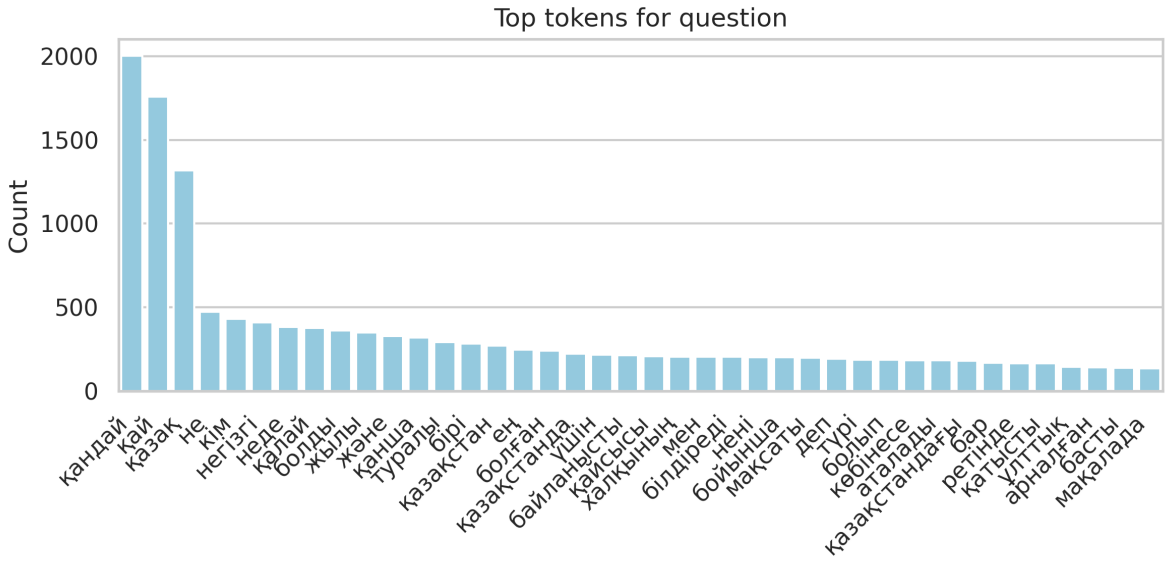

Figure 5: Most frequent tokens in question stems, highlighting common wh-terms and grammatical structures.

E Lexical Diversity by Category

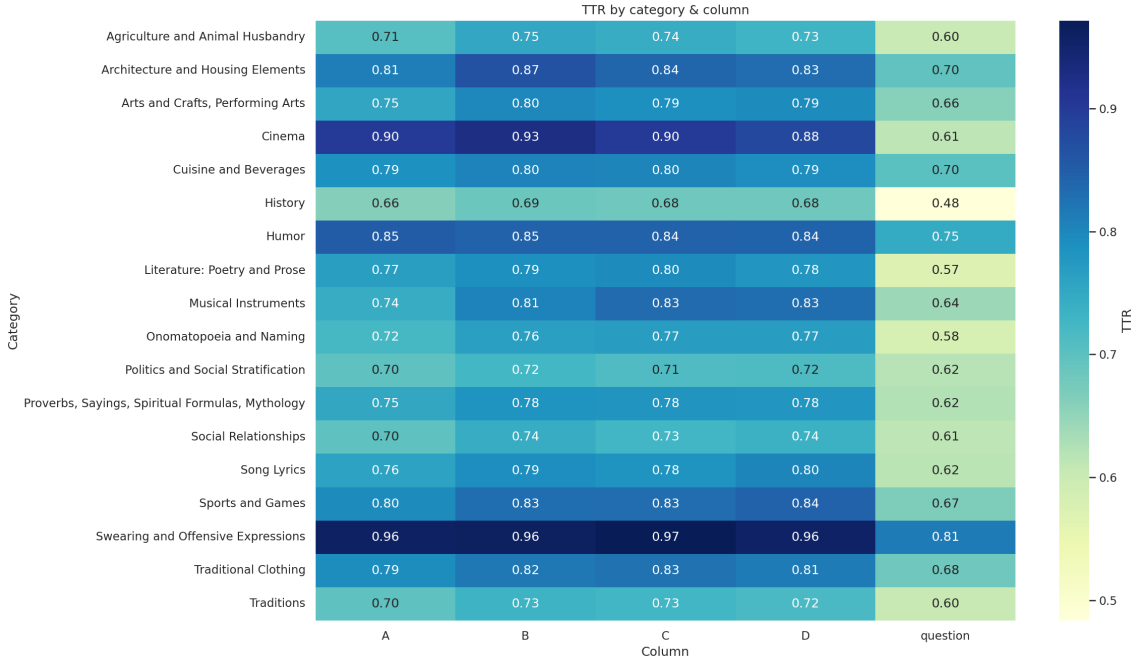

Figure 6: Type-token ratio (TTR) heatmap across categories, illustrating domain-specific variation in lexical richness.

F Answer Key Distribution

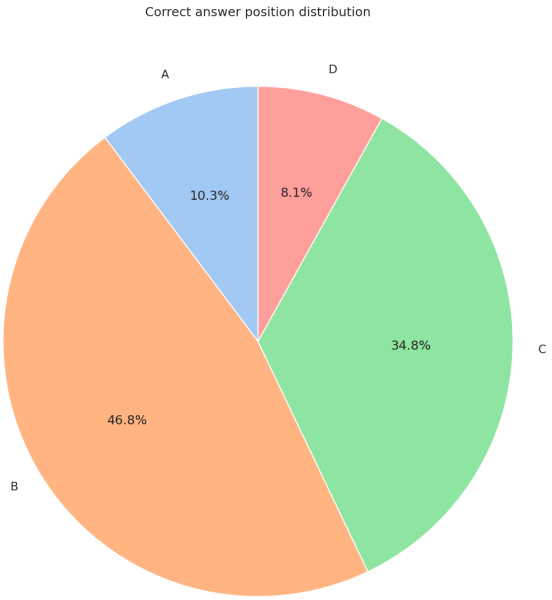

Figure 7: Distribution of correct answer positions (A–D), exposes bias in the dataset after human evaluation and fixes.

G Per-Category Model Accuracy

Model	Arch	Arts	Hist	Cinema	Cuisine	Lit	Swear	Instr	Onom	Polit	Proverb	Agric	Social	Sport	Song	Trad	Cloth	Humor	Avg
google/gemma-3-27b-it	0.759	0.708	0.706	0.688	0.748	0.669	0.649	0.671	0.667	0.788	0.725	0.765	0.746	0.651	0.737	0.779	0.714	0.740	0.722
meta-llama/Llama-3.3-70B-Instruct	0.747	0.714	0.709	0.625	0.696	0.655	0.684	0.675	0.657	0.817	0.703	0.741	0.741	0.675	0.708	0.729	0.723	0.663	0.709
meta-llama/Llama-3.1-70B-Instruct	0.719	0.703	0.697	0.625	0.724	0.653	0.667	0.650	0.652	0.810	0.707	0.737	0.744	0.679	0.685	0.713	0.737	0.673	0.703
nvidia/Llama-3_3-Nemotron-Super-49B-v1	0.723	0.686	0.691	0.598	0.671	0.641	0.649	0.689	0.633	0.792	0.677	0.741	0.741	0.656	0.683	0.715	0.710	0.694	0.694
inceptionai/Llama-3.1-Sherkala-8B-Chat	0.699	0.662	0.692	0.625	0.668	0.681	0.632	0.678	0.641	0.777	0.689	0.727	0.712	0.675	0.685	0.702	0.665	0.714	0.691
issai/LLama-3.1-KazLLM-1.0-70B	0.727	0.668	0.703	0.571	0.675	0.657	0.702	0.636	0.622	0.773	0.693	0.725	0.714	0.665	0.687	0.708	0.696	0.684	0.689
google/gemma-3-12b-it	0.731	0.673	0.669	0.661	0.664	0.647	0.544	0.657	0.630	0.737	0.709	0.694	0.680	0.623	0.693	0.713	0.719	0.679	0.679
mistralai/Mistral-Small-24B-Instruct-2501	0.687	0.681	0.685	0.589	0.671	0.625	0.684	0.661	0.634	0.724	0.659	0.712	0.697	0.618	0.668	0.715	0.692	0.704	0.676
Qwen/Qwen2.5-32B-Instruct	0.699	0.614	0.604	0.598	0.570	0.649	0.509	0.618	0.612	0.717	0.619	0.642	0.638	0.608	0.643	0.658	0.643	0.694	0.633
google/gemma-3-12b-pt	0.671	0.611	0.589	0.518	0.629	0.561	0.649	0.594	0.531	0.695	0.665	0.665	0.675	0.561	0.637	0.692	0.688	0.643	0.624
Qwen/QwQ-32B	0.651	0.605	0.601	0.589	0.573	0.637	0.526	0.590	0.597	0.658	0.615	0.639	0.645	0.599	0.599	0.622	0.589	0.699	0.617
deepseek-ai/DeepSeek-R1-Distill-Llama-70B	0.699	0.576	0.603	0.464	0.587	0.565	0.632	0.565	0.504	0.667	0.625	0.639	0.643	0.561	0.601	0.620	0.634	0.638	0.602
Qwen/Qwen2.5-14B-Instruct	0.631	0.627	0.573	0.536	0.601	0.605	0.456	0.601	0.572	0.658	0.561	0.613	0.638	0.599	0.585	0.649	0.580	0.622	0.600
deepseek-ai/DeepSeek-R1-Distill-Qwen-32B	0.635	0.614	0.583	0.589	0.577	0.629	0.491	0.565	0.576	0.634	0.579	0.634	0.589	0.547	0.589	0.640	0.598	0.633	0.600
google/gemma-3-4b-pt	0.643	0.616	0.579	0.509	0.584	0.543	0.509	0.544	0.536	0.631	0.595	0.613	0.601	0.613	0.578	0.608	0.563	0.602	0.585
google/gemma-3-4b-it	0.618	0.578	0.575	0.438	0.580	0.557	0.544	0.640	0.548	0.636	0.581	0.615	0.589	0.552	0.572	0.576	0.643	0.566	0.583
meta-llama/Llama-3.1-8B-Instruct	0.647	0.614	0.573	0.482	0.535	0.545	0.614	0.530	0.507	0.600	0.589	0.606	0.618	0.561	0.570	0.576	0.580	0.622	0.575
issai/LLama-3.1-KazLLM-1.0-8B	0.598	0.568	0.576	0.455	0.549	0.511	0.649	0.516	0.462	0.638	0.569	0.611	0.628	0.524	0.557	0.597	0.585	0.602	0.566
nvidia/Llama-3.1-Nemotron-Nano-8B-v1	0.341	0.351	0.359	0.339	0.374	0.311	0.386	0.392	0.327	0.355	0.365	0.370	0.340	0.406	0.347	0.346	0.402	0.342	0.354
TilQazyna/llama-kaz-instruct-8B-1	0.233	0.235	0.282	0.295	0.318	0.281	0.193	0.325	0.264	0.291	0.327	0.287	0.249	0.288	0.261	0.264	0.237	0.265	0.277

Table 6: Per-category accuracy (and macro average) for each evaluated model. Column abbreviations: **Arch** = Architecture/Housing, **Arts** = Arts/Crafts, **Lit** = Literature, **Swear** = Swearing expressions, **Instr** = Musical instruments, **Onom** = Onomatopoeia, **Polit** = Politics/Social, **Proverb** = Proverbs & Mythology, **Agric** = Agriculture, **Trad** = Traditions, **Cloth** = Traditional clothing.

E Semi-Automated Data Generation Pipeline

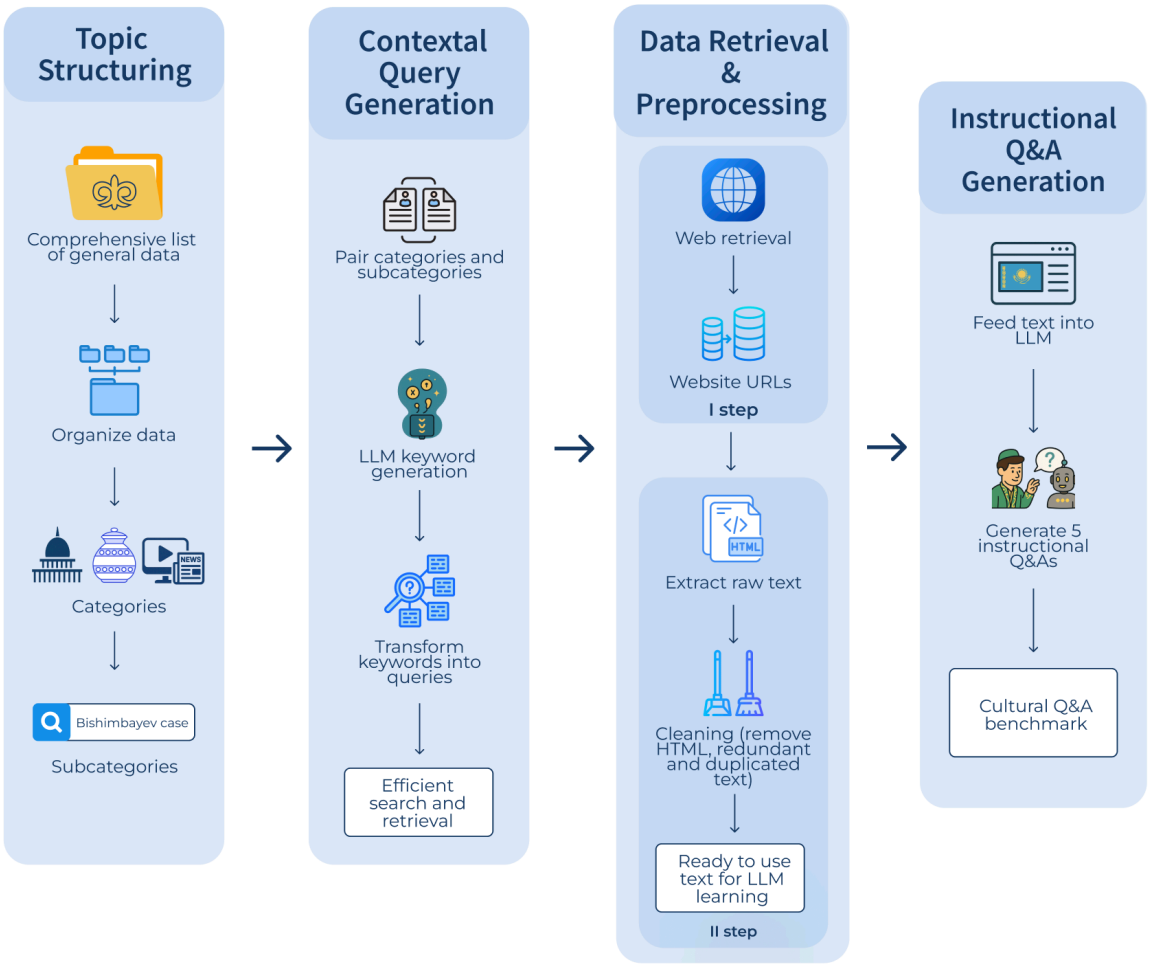

Figure 8: Overview of the semi-automated pipeline used to generate culturally-grounded instructional Q&A benchmark.